Encodage d'activité neuronale à partir de réseaux LSTM

Antonin Verdier

Paris, 18 juin - 21 septembre 2018

Neurospin - Département Unicog

Projet Le Petit Prince

Encadrants responsables : Yair Lakretz, Christophe Pallier

Responsable du Laboratoire : Stanislas Dehaene

École Normale Supérieure

Table des matières

1 Présentation du laboratoire

NeuroSpin, dirigé par Stanislas Dehaene, est un centre de recherche pour l'innovation en imagerie cérébrale. Les travaux qui y sont menés s'inscrivent dans les deux axes imagerie biomédicale et innovation diagnostique et thérapeutique.

Au département NeuroSpin, physiciens, mathématiciens et neuroscientifiques s'allient pour développer en synergie les outils et les modèles qui permettront de mieux comprendre le fonctionnement du cerveau normal et pathologique, avant ou après traitement. Centrées sur la neuroimagerie, les recherches conduites sont de plusieurs natures : développement technologiques et méthodologiques (acquisition et traitement des données), neurosciences cognitives, neurosciences précliniques et cliniques.

NeuroSpin comporte 5 entités de recherche : unité d'imagerie par résonance magnétique à très haut champ et de spectroscopie ; unité d'analyse et traitement de l'information, unité de neurosciences cognitives (UNICOG), une unité mixte de recherche, sous tutelle du centre de l'énergie atomique, Université Paris-Sud et Inserm ; unité neuro-imagerie applicative clinique et translationnelle, rattachée à l'UMR 1129.

L'équipement à Neurospin est exceptionnel. On y trouve les appareils à IRM les plus puissants au monde comme des IRM cliniques 3T, 7T et bientôt 11,7T, des machine à IRM précliniques pour le petit animal 7T, 11,7T et 17T. C'est également un centre d'électro-encéphalographie et de magnéto-encéphalographie. Prochainement, un microscope tri-photonique sera mis en place.

J'ai effectué mon stage au département UNICOG. Sa mission est d'étudier les bases cérébrales des fonctions cognitives, chez l'homme sain et chez des patients atteints de maladie neurologique, en développant et en exploitant les méthodes de neuro-imagerie conjointement à l'utilisation de paradigmes expérimentaux issus de la psychologie cognitive. Les outils employés sont l'imagerie par résonance magnétique fonctionnelle (IRMf) ainsi que la magnéto- et électroencéphalographie. C'est l'équipe Neuroimagerie du Langage dirigée par Christophe Pallier qui m'a accueilli.

2 Introduction

La perception du langage par le cerveau humain est un phénomène complexe qui met en jeu de nombreuses aires cérébrales interagissant constamment [3]. La modélisation d'un tel système linguistique représente un enjeu majeur dans sa compréhension. Dans un modèle, l'information peut être captée, déformée et enlevée à tout moment. L'étude de ce modèle pourra apporter un nouveau regard sur la manière dont nous apprenons les mots et parallèlement sur la manière dont les troubles du langage naissent.

Dans ce mémoire, il est question de mettre en corrélation deux visions : biologique et informatique. En présentant une même histoire à un sujet sain et à un programme d'apprentissage neuronal il devient possible de prédire le comportement de l'un grâce au comportement de l'autre. En guise de premier concept, nous avons tenté de prédire l'activité des neurones humains à partir de l'activité des neurones informatiques. Pour ce faire, il est nécessaire dans un premier temps de définir un modèle informatique sachant traiter le langage naturel.

La fin des années 1980 fut témoin de l'apparition des algorithmes de Machine Learning dans le traitement du langage naturel. Une conséquence quasi-direct de la première loi de Moore. Les techniques basées sur des règles syntaxiques développées par Noam Chomsky en 1957 [4] laissèrent place à une approche plus probabiliste. Les modèles linguistiques en cache, un sous type de modèle probabiliste, sont d'ailleurs à la base de la plupart des programmes de reconnaissance vocale actuels [12]. Aujourd'hui, les algorithmes de réseaux de neurones à apprentissage supervisé, semi-supervisé ou non-supervisé dominent le domaine. C'est dans ce contexte que nous avons choisi de travailler avec un réseau Long Short-Term Memory (LSTM).

En effet, les réseaux LSTM semblent être des candidats de choix puisque leur structure même vise à conserver la mémoire des entrées précédentes [16]. Se souvenir des mots présents au début d'une phrase est en effet primordial pour en comprendre le sens une fois sa lecture terminée. Le développement d'un tel outil fut réalisé en partenariat avec le laboratoire d'Intelligence Artificielle de Facebook. Après entraînement du réseau (qui peut s'assimiler à l'apprentissage d'une langue) une version texte de "Le Petit Prince" d'Antoine de Saint-Exupéry lui est présenté. Nous disposons donc à ce point des données "informatiques" à savoir l'activation des neurones du réseau pour une certaine histoire.

Les données biologiques d'activité cérébrale furent récupérées par imagerie par résonance magnétique fonctionnelle (IRMf). C'est en effet une méthode non radiative très souvent utilisé en neuroimagerie puisque rend extrêmement bien compte des tissus mous et de leur activité. En présentant la version audio du texte "Le Petit Prince" d'Antoine de Saint-Exupéry à ces volontaires, il est alors possible d'enregistrer indirectement la répartition et l'évolution de leur activité cérébrale, et ce notamment grâce au signal blood-oxygen-level dependant (BOLD)[14]. C'est en présentant le même texte au sujet sain et au programme informatique qu'il devient possible de corréler le comportement de ces derniers.

Ce mémoire s'inscrit dans le projet "Le Petit Prince", projet de neuroimagerie de grande envergure. Il est en lien direct avec le projet Neuroparsing de l'Université de Cornell. L'objectif premier de ce stage est la création d'un programme de prédiction permettant de prédire l'activité cérébrale du sujet sain à partir des activation du réseau LSTM. Le postulat de ce modèle de prédiction est qu'une prédiction juste montre

la présence de l'information (en tant que concept mathématique) dans le réseau de neurones LSTM. La force de cette approche informatique est l'absence quasi-totale de subjectivité humaine dans le choix des paramètres du modèle et de son contenu. Le présent rapport rend compte de la création de ce modèle de prédiction.

3 Matériels et Méthodes

3.1 Introduction de concepts

3.1.1 Word Embeddings

Une manière de se représenter mathématiquement une phrase ou un mot passe par la notion de vecteur sémantique [11]. Chaque mot est représentable dans un espace continu de très haute dimension. L'étude de la distribution des différents mots dans un texte fourni des indices sur leurs caractéristiques communes. Deux mots sémantiquement proches sont susceptibles d'apparaître dans le même contexte. Le principe de continuité est fondamental et autorise des calculs simples tel que celui de la distance euclidienne entre deux vecteurs, souvent représenté comme la "distance sémantique" entre ces deux mots. La figure 1 illustre quelques unes de ces représentations :

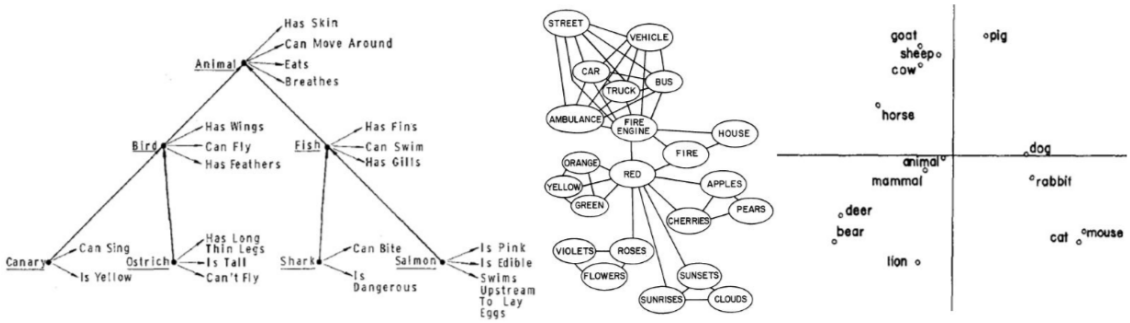


FIGURE 1 – Gauche : Réseau sémantique représentant les relations entre concepts et propriétés via des noeuds et des connexions. La distance sémantique est schématisée par le nombre de nœuds et connexions entre deux concepts. [6] Centre : Réseau sémantique représentant les relations comme des activations répandues. Bien que la distance sémantique soit mesurée similairement à la figure de gauche, ce modèle donne une information sur la similarité sémantique par le nombre de connexions [5]. Droite : Les concepts ou les mots sont représentés dans un réseau défini par deux axes, ici en fonction de la taille (X) et de la dangerosité (Y) [15].

Ces approches permettent de garder les informations relationnelles entre les mots puisque ces derniers sont placés dans un espace continu. Cette approche se fonde sur une hypothèse distributionnelle. L'idée initiale est qu'il existe une corrélation entre similarité de contextes et similarités de sens. Ainsi, en étudiant le contexte relatif à un mot, il est possible d'approcher son sens. Cette méthode est en générale empirique puisqu'appuyée sur des données réelles pour extraire le sens. Enfin, ce modèle repose aussi sur la théorie structuraliste du langage dans laquelle le sens d'un mot est entièrement modélisable par des relations syntagmatiques et paradigmatiques.

3.1.2 Méthodes de régression

La régression des moindres carrés est l'une des régressions les plus simples et les plus utilisées. L'objectif est de minimiser le carré de la différence entre l'estimation et la valeur vraie. En considérant une matrice X de dimensions $mesures \times features$, une matrice W de paramètres à estimer et enfin une matrice Y de mesures :

$$W^* = argmin(||XW - Y||^2) \tag{1}$$

Où W^* désigne la matrice W optimale.

Cependant, notre modèle comporte beaucoup de *features*. Les *features* sont les composantes choisies pour représenter les données. Par exemple, pour prédire si il pleuvra le lendemain, il est possible de se baser sur la position des nuages. Dans ce cas, la position des nuages est une *feature*. Un nombre important de *features* favorise l'overfitting [8]. L'overfitting se définit comme un modèle trop performant sur l'échantillon d'entraînement et donc très difficilement extrapolable (figure 2).

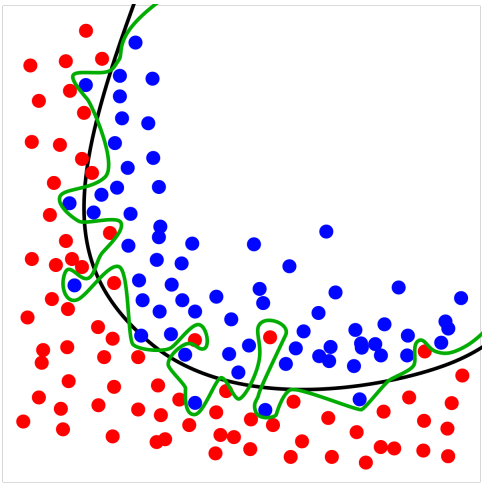


FIGURE 2 – Représentation à 2 dimensions du phénomène d'overfitting

La courbe verte représente un modèle "overfitté", très performant sur l'échantillon. Cependant, intuitivement le modèle représenté par la courbe noire semble bien plus approprié. L'overfitting est ainsi une conséquence du nombre important de *features*. Une solution consiste à l'ajout d'un terme régulateur visant à diminuer la valeurs des coefficients de X. En forçant des valeurs modestes, le modèle est alors moins complexe et cela limite l'overfitting.

$$W^* = argmin(||XW - Y||^2 + \alpha||W||^2) \tag{2}$$

Cette formule de régression correspond à la régression de Ridge ou régression de Tikhonov. α représente l'amplitude de la pénalisation des coefficients. Cet hyperparamètre contrôle donc directement la complexité du modèle, un alpha élevé forcera des valeurs de paramètres W faibles donc un modèle simple et inversement pour un alpha faible. L'influence des variations de alpha est illustré dans la figure 3.

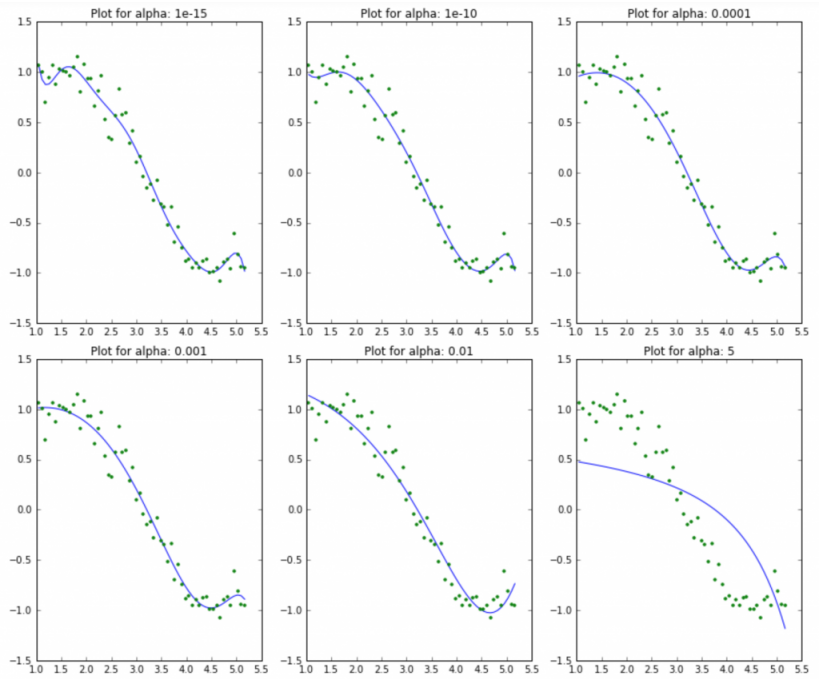


FIGURE 3 – Régression de Ridge selon α. Une valeur de alpha trop petite est inutile et ne contrebalance pas la tendance du modèle à l'overfitting. En revanche, une valeur trop élevée donne un modèle trop simple absolument non-représentatif des données.

3.1.3 Validation croisée

La validation croisée est une technique visant à mieux estimer les performances d'un modèle et à les rendre plus extrapolables. Pour cela, l'ensemble des données est scindé en un certain nombre de blocs. Une partie de ces blocs va être conservée comme un ensemble dit d'entraînement. La seconde partie sera considérée comme un ensemble test. Il est d'usage d'allouer 80% des données à l'entraînement et 20% au test. Pour une base de données scindées en 10 blocs, 8 seront donc réservés à l'entraînement et 2 au test. La validation croisée consiste à entraîner non pas un seul mais bien 10 modèles différents avec à chaque fois une configuration différente des groupes d'entraînement et de test. En moyennant l'erreur sur chaque itération, on obtient alors une performance globale du modèle sur l'ensemble des données disponibles. Ainsi on optimise notre usage des données disponibles et de ce fait l'extrapolation des performances de notre modèle. La figure 4 illustre schématiquement le fonctionnement d'une telle validation croisée.

3.2 Données IRMf

3.2.1 Acquisition

L'acquisition des données cérébrales fut réalisée en présentant une version audio anglaise de "Le Petit Prince" d'Antoine de Saint-Exupéry à des sujets natifs alors que l'activation cérébrale était enregistrée via un IRMf Phillips 3 Tesla.

Le projet global portait sur deux langues, le français et l'anglais. A ce jour, seul les données anglaises sont traitées autant pour la partie informatique que pour les mesures BOLD. Ceci est principalement dû aux commodités linguistiques offertes par

FIGURE 4 – Principe schématique d'une validation croisée

la langue anglaise, le français présentant plus de caractères spéciaux et de ponctuation. Les données IRMf prétraitées des 51 participants anglophones m'ont été fournies par nos collaborateurs du projet Neurosparsing à l'Université de Cornell.

Pour rendre l'écoute du livre audio plus agréable pour les participants et favoriser leur concentration, le texte audio du livre en anglais a été divisé par nos collègues à l'Université de Cornell en neuf blocs tenant compte de la thématique et de la durée, de sorte qu'aucun bloc ne dépasse 15 minutes. Une seconde de silence a été ajoutée entre tous les blocs. A la fin de chaque bloc, des questions de compréhension sont

Bloc	Chapitres	Durée
1	1-3	09 :31
2	4-6	10 :04
3	7-9	11 :27
4	10-12	10 :14
5	13-14	08 :58
6	15-19	11 :34
7	20-22	10 :58
8	23-25	09 :52
9	26-27	12 :23

TABLE 1 – Durée des blocs d'écoute

posées au sujet, afin de vérifier son attention pendant l'écoute. L'intégration des sujets dans cette étude dépendait de deux facteurs. Les sujets devaient être droitiers selon le test de latéralité d'Edimbourg, afin de favoriser la reproductibilité de l'expérience. De plus, ils ne devaient pas avoir lu ou entendu parlé de "Le Petit Prince" les 5 dernières années, encore une fois pour des questions de reproductibilité mais aussi pour éviter l'activation intempestive des zones mémorielles.

3.2.2 Pré-traitement

Les coupes d'IRM ont été acquises grâce à une séquence multi-echo EPI spécialement conçue pour optimiser le rapport signal sur bruit de la réponse BOLD[13]. Il est nécessaire de pré-traiter les données, il faut alors utiliser un script spécifique (me-ica : https ://bitbucket.org/prantikk/me-ica). Le pré-traitement des données anglaises a été fait à l'Université de Cornell.

Suite à cela, nous obtenons des fichiers d'extension .nii contenant pour chaque voxel 3 valeurs correspondant aux dimensions spatiales et 1 valeur temporelle. De plus, à chaque fichier est associé un ensemble de métadonnées. Le volume cérebral était ainsi découpé en voxels de dimension 2.0x2.0x2.0mm.

3.3 LSTM

3.3.1 Introduction

Les réseaux Long Short-Term Memory (LSTM) font partie d'une catégorie plus large d'algorithmes dits de réseaux de neurones. De tels réseaux se constituent d'un élément de base, le neurone formel. Un neurone formel est une entité mathématique tripartite : entrées, fonction d'activation et sortie. A un neurone formel est présenté plusieurs entités, dans notre cas des scalaires. Chaque scalaire i est pondéré par un poids W. Pour déterminer la valeur entrante dans le neurone on effetue la somme de chaque entrée pondérée. Cette valeur entrante est ensuite fournie à la fonction d'activation (qui prend souvent la forme d'une tangente hyperbolique) créant ainsi la sortie o. Cette dernière peut constituer la réponse finale au problème ou bien servir à son tour comme valeur entrante dans un prochain neurone.

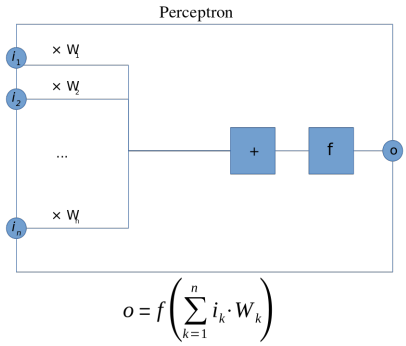
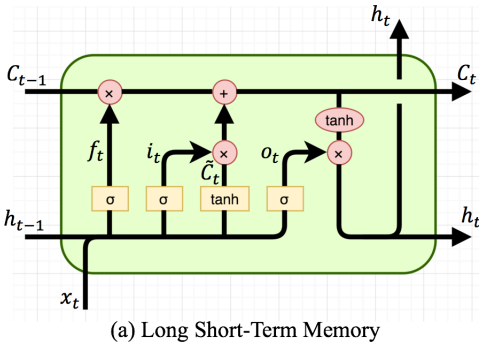


FIGURE 5 – Schéma d'un Perceptron. k correspond au nombre d'entrées



(a) Long Short-Term Memory

FIGURE 6 – Schéma d'un neurone d'un réseau Long Short-Term Memory

Un réseau de neurone est ainsi un assemblage de neurones formels organisés en couches. Chaque neurone d'une couche reçoit l'ensemble des sorties de la couche précédente. Un exemple de réseau utilisant les neurones formels est le perceptron (figure 5)

Le réseau neuronal utilisé ici appartient à une catégorie plus complexe, les Long Short-Term Memory (LSTM). Ce type de réseau est particulièrement performant dans les tâches nécessitant de se "souvenir" des précédentes entrées. Cette particularité prend son sens dans l'étude du langage où il paraît en effet nécessaire de se souvenir du

début d'une phrase afin d'en extraire le sens. Pour se doter de cette caractéristique, un neurone de LSTM n'est pas seulement représenté par une fonction d'activation mais par un système de trois portes : *forget gate, input gate* et *sigmoid gate*.

Sur le schéma ci-dessus (figure 6), il est important de distinguer le "Cell state" (CS) et le "Hidden State" (HS), deux caractéristiques fondamentales pour le fonctionnement d'un neurone LSTM (là où un neurone formel n'en possède qu'un). Le CS, noté C_t sur la figure est quasi-spécifique de ce type de réseau, à savoir un flux d'information qui "court" tout le long du réseau. Le HS, noté h_t est le flux d'information d'un neurone et peut être comparé à l'information d'un neurone formel (moyennant les opérations supplémentaires).

La *forget gate*, f_t, comme son nom l'indique est dévouée à la conservation ou à l'élimination de l'information. L'*input gate*, i_t, est elle responsable de l'information nouvelle à ajouter au CS. Enfin, la *sigmoid gate*, o_t, se charge de l'information de sortie. La combinaison de ces trois portes (représentées sur la figure 2) est l'essence de ces réseaux.

3.3.2 Structure du réseau

Le réseau de neurone LSTM utilisé pour modéliser le processus de traitement du langage est une structure à cinq couches comme présenté sur la figure 7. La première et la dernière couche comportent 13000 neurones. Les couches intermédiaires ont 650 neurones.

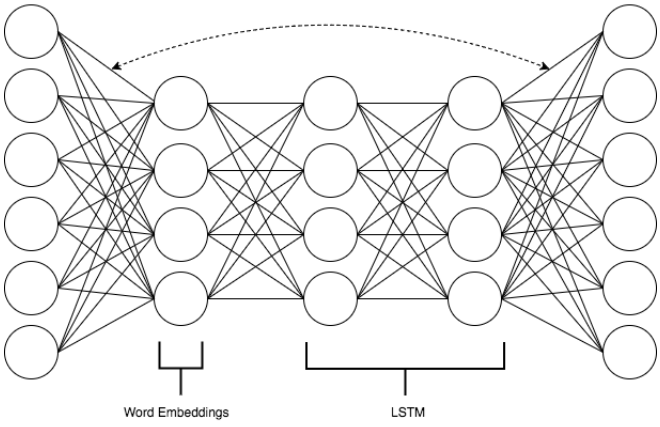


FIGURE 7 – Schéma du réseau LSTM

La première couche (à gauche) correspond à l'entrée du réseau. Il y est injecté un vecteur dit *one hot* de 13000 lignes. La valeur "1" correspond à un mot parmi les 13000 du corpus de vocabulaire. Cette entrée à 13000 dimensions est ensuite réduite dans un espace de seulement 650 dimensions, plus dense. Cette couche du réseau neuronal s'apparente aux Word Embeddings. Habituellement, ces derniers sont générés selon des critères humains définis, sémantiques ou grammaticaux comme par exemple la valeur de beauté du mot, sa connotation au vivant, à l'écriture etc. Dans notre cas, ces 650 critères de définition d'un mot sont purement abstraits et ne sont donc sujets à aucun

biais humain ou textuel. Cette couche est constituée simplement de neurones formels. Les deux couches suivantes sont des couches formées de neurones dits LSTM (figure 6). Suite aux couches intermédiaires, l'information est à nouveau projetée dans un espace à 13000 dimensions. Les poids (ou pondération) entre l'avant dernière et la dernière couche sont forcés à être identiques aux poids entre la première et la seconde couche. Cela assure une cohérence dans l'apprentissage et dans le changement de nombre de dimensions. Il est nécessaire de décompresser les dimensions symétriquement à la façon dont on les a compressées

3.3.3 Extraction des activations

La structure du réseau de neurones étant définie, ce dernier peut à présent être entrainé sur des données. Afin de lui apprendre la langue anglaise, le réseau LSTM fut entrainé sur un corpus de plus de 1 million de textes catégorisés provenant de Wikipédia. À chaque itération, le réseau ajuste ses paramètres internes. Au terme de cette apprentissage, le réseau possède des paramètres internes optimaux et est apte à réaliser des prédictions.

Une fois entrainé, il est présenté au réseau le texte de "Le Petit Prince". La version texte modifié du livre (pour qu'elle corresponde parfaitement au livre audio) a été découpée en 9 blocs. Chaque bloc fut scindé en tokens. Un token peut être soit un mot soit une ponctuation. Chaque token est alors présenté individuellement et dans l'ordre au réseau LSTM. Pour chacun, les activations HS des couches LSTM furent récupérées et stockées dans un fichier.

3.4 Model de prédiction

L'objet de mon stage fut la création de ce programme de prédiction. Il s'agit actuellement d'un programme informatique Python s'appuyant sur les données IRMf récoltés à l'Université de Cornell selon les modalités décrites chapitre 2.1 et sur les données informatiques de réseau LSTM décrites chapitre 3.3.2.

3.4.1 Création des régresseurs

Il existe une différence physique fondamentale entre les activations du réseau LSTM et le signal BOLD que le modèle tente de prédire. En effet, alors que les activations sont discrètes et instantanées, le signal BOLD de l'IRMf est une manifestation indirect d'un phénomène biologique. Il est de ce fait continu et variable. Pour palier à cette différence, les activations doivent être traitées pour se comporter comme une fonction continue évoluant au cours du temps. À partir de l'amplitude des activations du réseaux LSTM et des offsets des mots correspondants (le temps auquel le mot a fini d'être prononcé sur le livre audio), il est possible de constituer en quelque sorte une fonction BOLD du réseau. Cette opération est réalisée par convolution de ces deux paramètres avec la fonction de réponse hémodynamique (HRF). Ce script simule la réponse BOLD qu'aurait produit l'activation du neurone informatique si celui ci était un neurone biologique.

3.4.2 Validation croisée hierarchisée

Les chapitres 3.1.2 et 3.1.3 ont montré l'importance, respectivement, du paramètre α de la régression de Ridge et de la validation croisée. Postulant que chaque voxel est une zone se comportant différement de la zone voisine, il paraît adéquat de considérer non pas un modèle de prédiction global mais bien un modèle spécifique pour chaque voxel. Au vu de l'importance d'alpha dans la régression de Ridge, il convient de l'optimiser pour chaque voxel.

Une solution à cela consiste à réaliser une validation croisée hierarchisée. Les huit blocs d'entrainement sont à nouveau divisés de manière à dégager un groupe d'entraînement niché de sept blocs et un groupe de validation de 1 bloc. En fournissant une liste de valeurs vraisemblables d'alpha(s), l'algorithme de validation croisée va permettre de choisir le meilleur alpha pour chaque voxel. Ces alphas sont ensuite réinjectés dans la régression (voir Chapitre 3.4.3).

3.4.3 Intégration globale

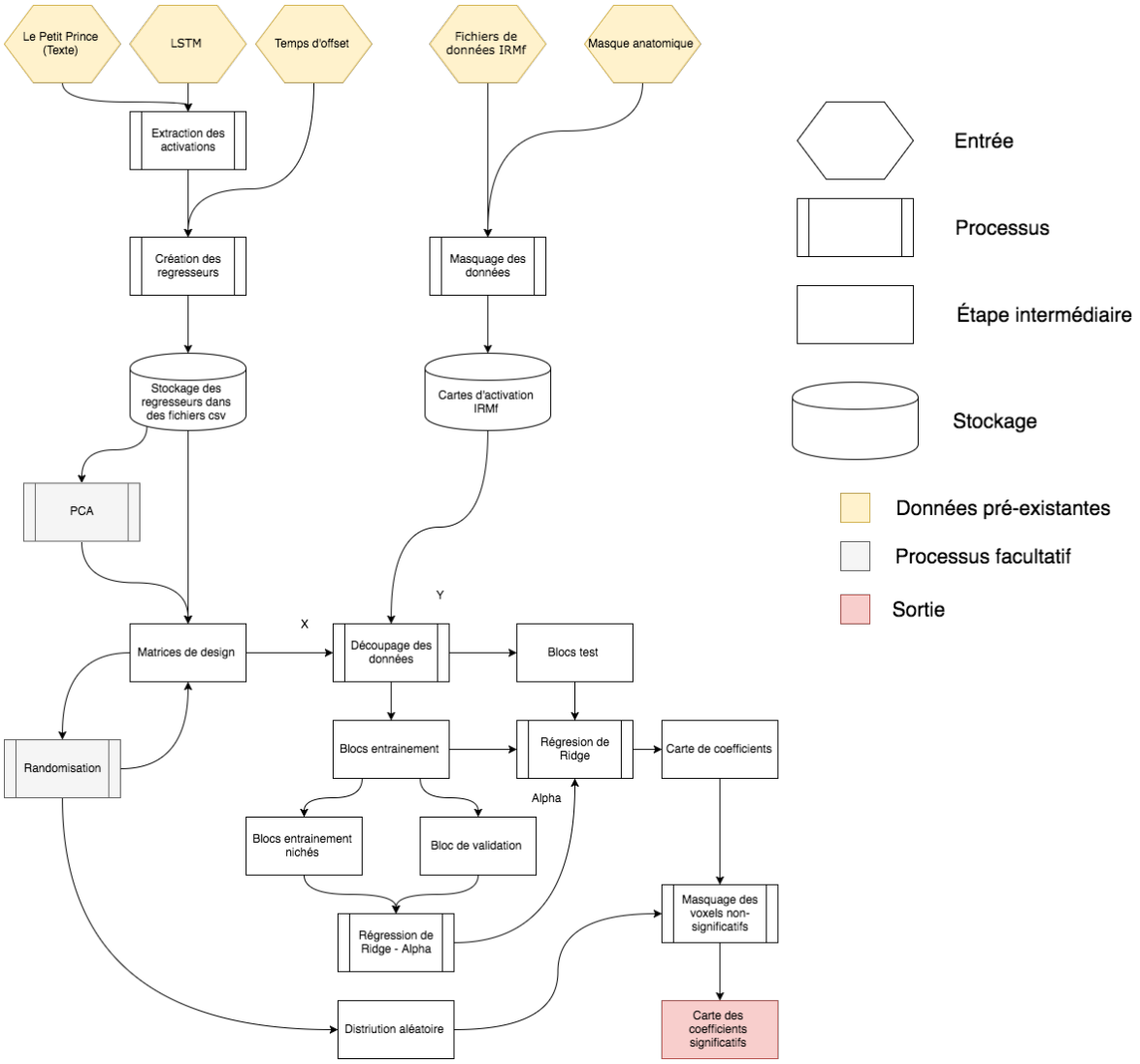


FIGURE 8 – Schéma global du programme de prédiction

Le programme commence par un traitement des données IRMf afin de les préparer à la régression. Au départ, un voxel donné est caractérisé par 3 grandeurs spatiales décrivant sa position dans le cerveau pour un scan donné. Les scans bout à bout constituent une évolution temporelle discrète de 2 secondes en 2 secondes. Pour faciliter la régression, les voxels sont étalés dans un vecteur à une seule dimension. Les données IRMf d'un bloc sont donc une matrice dont une ligne représente un scan du bloc et les colonnes les voxels. Après masquage des données IRMf, le volume cérébral était représenté par 219486 voxels.

Parallèlement, les activations issues du réseau neuronal sont placés dans une matrice dont les lignes représentent les scans et les colonnes, les *features*. Ici, les *features* sont variables et leur choix détermine le type de modèle. Les *features* représentent les données utilisées pour prédire l'activation cérébrale. Par exemple, il est possible d'essayer de prédire l'activité cérébrale avec seulement le nombre de lettres de chaque mot en guise de donnée d'entrée. Un des objectifs de ce modèle est aussi de comparer les performances des prédictions en fonction des données choisies au départ. Nous avons choisi d'entraîner trois grands types de modèles, Basic Features (BF), Basic Word Embedding (BFWE) et Basic LSTM (BFLSTM). Ces modèles diffèrent par les *features* utilisées au départ.

Pour le modèle BF, chaque voxel possède 5 *features* : RMS, Frequency, F0, Bottom Up et Wordrate. RMS pour pression acoustique efficace, un paramètre acoustique qui peut s'apparenter au volume du signal sonore. Frequency est un paramètre indiquant la fréquence d'apparition du token en question dans le corpus référence. F0 correspond à la fréquence fondamentale du son du livre audio présenté au sujet pendant l'écoute dans l'IRMf. Wordrate peut s'apparenter à une indication sur la vitesse à laquelle le mot est prononcé dans le livre audio. Enfin, Bottom Up fait référence au "Bottom Up parsing" une technique mettant en lumière la structure grammaticale d'une phrase. Le modèle BF se décrit comme une base assez acoustique, qui permet de reproduire les conditions d'écoute alors que les données présentées au réseau de neurone sont textuelles.

Le modèle BFWE est composé de 305 *features* exactement, 5 étant identiques au modèles BF. Les 300 autres correspondent chacune à une dimension des Word Embedding des mots énoncés pendant le scan, comme expliqué chapitre 3.1.1. Les Word Embeddings utilisés ici ont été générés préalablement selon la méthode GloVe. Ce modèle rend compte de l'aspect sémantique couplé aux informations auditives.

Enfin, le modèle BFLSTM est composé de 1305 *features*, 5 étant identiques au modèle BF. Les 1300 autres correspondent aux activations HS des deux couches LSTM du réseau LSTM. Ce modèle est le modèle d'intérêt. Les Word Embeddings étant générés automatiquement par le réseau à la couche précédante, l'information encodée se veut de plus haut niveau mais également beaucoup plus abstraite.

Ces trois modèles sont indépendants et correspondent chacun à un lancement du programme de prédiction.

Une fois les jeux de données prêts, le programme peut alors entrer dans la phase d'apprentissage. Les neufs blocs sont subdivisés en huit blocs d'entraînement et un bloc de test. Ces blocs d'entraînement sont alors découpés à nouveau en sept blocs d'entraînement et un bloc de validation. Sur le principe de la validation croisée hierarchisée décrite plus haut, les sept et un blocs permettent de trouver une valeur de l'hyperparamètre α optimale pour chaque voxel indépendamment. Cette liste de valeurs d'alpha

est ensuite réinjecté comme paramètre dans la fonction Ridge de premier niveau. S'en suit une validation croisée de huit et un bloc. Chaque phase de test (lorsque le modèle tente de prédire les mesures à partir d'un échantillon jamais rencontré) génère un vecteur de coefficient de détermination (voir chapitre 3.5). Ce vecteur est ensuite nettoyé. En effet, les valeurs négatives sont ramenées à zéro (principalement pour des questions de génération de figures) et les valeurs supérieures à 0.99 sont également ramenées à zéro. Cette dernière mesure cherche à éliminer les artéfacts de régression, les valeurs "trop justes pour être vraies".

Enfin, les voxels non-significatifs sont retirés selon la méthode statistique explicitée Chapitre 2.4. Il en ressort un vecteur de coefficients de détermination qui, fournit en entrée de la fonction de génération de carte, est projeté sur l'espace cérébral et est utilisé pour générer les figures (voir Résultats).

3.5 Statistiques

Le coefficient de détermination R^2 permet de juger de la qualité d'une régression.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y}_i)^2} \tag{3}$$

Où n est le nombre de mesures, y_i la valeur de la mesure i, $\hat{y}_i$ la valeur prédite correspondante et $\bar{y}_i$ la moyenne des mesures. Ce coefficient est calculé pour chaque voxel.

Parmi les activations prédites, certaines ne sont pas significatives. Pour corriger cela, il est nécessaire de développer un filtre basé sur une distribution nulle.

Le programme est ainsi lancé avec un paramètre impliquant une randomisation des colonnes de la matrice des données LSTM d'entrainement. En les mélangeant, les *features* d'entrainement ne sont ainsi plus alignées avec les *features* de test. Il en résulte des coefficients de détermination uniquement dûs au bruit et à la structure du modèle. 1008 simulations de ce type ont été réalisées. Une matrice est ensuite créée, pour chaque sujet et chaque type de modèle (BF, BFWE, BFLSTM) avec, pour nombre de lignes le nombre de simulations aléatoires réalisées et sur chaque ligne les coefficients de détermination. Dans un second temps, cette matrice est transformée en vecteur colonne. Pour chaque simulation aléatoire, le coefficient de détermination le plus grand est sélectionné parmi les 219486 voxels. Il en résulte un vecteur de 1008 maxima de coefficients de détermination.

Pour qu'un voxel soit significatif lors d'une simulation, la valeur de son coefficient de détermination est comparé à ce vecteur de maximums. La p-value d'un voxel correspond au nombre de maximum de coefficients de détermination aléatoire dont la valeur est plus importante que la valeur du coefficient de détermination de la simulation à filtrer, le tout divisé par le nombre de simulations aléatoires réalisées. Cette opération est effectuée pour chaque voxel. Il en résulte un vecteur binaire où "1" correspond à une valeur de p-value inférieure au seuil défini (dans notre cas 0.001) et "0" à une valeur supérieure. Ce vecteur est par la suite multiplié au vecteur de coefficients de détermination élément par élément. Toutes les valeurs égale à 0 sont retirées et seules les valeurs non nulles sont représentées sur les figures.

4 Résultats

4.1 Glassbrain

Les figures n'ont été générées que pour un seul sujet pour des raisons de temps de calcul. La génération préliminaire d'autres figures suggère toutefois des résultats similaires pour les autres sujets.

Le coefficient de détermination peut s'assimiler à la capacité de prédiction des variances des mesures par le modèle. Un coefficient de 1 est ainsi synonyme d'un modèle parfait. Une meilleure prédiction dans une zone du cerveau montre que les données du réseau LSTM contiennent plus d'informations en lien avec le rôle de cette zone.

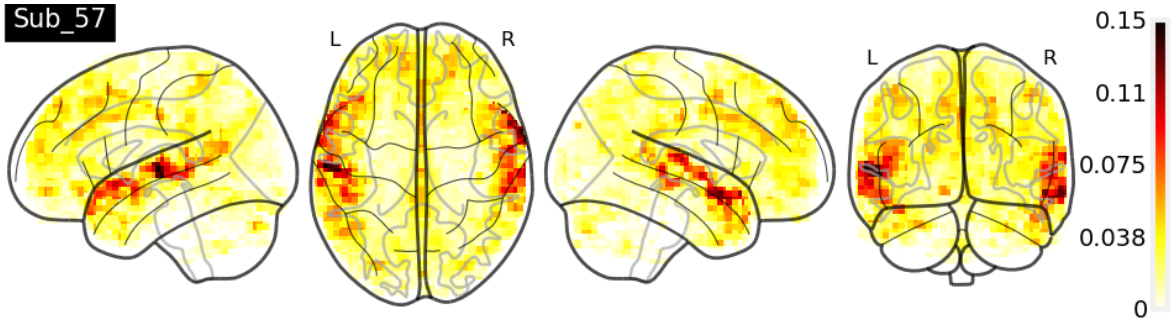


FIGURE 9 – Coefficient de détermination BF

Les régions cérébrales les plus rouges correspondent aux régions les mieux prédites par le modèle. Sur la figure 9, on distingue une meilleur prédiction au niveau du gyrus temporal supérieur, cortex auditif primaire (partie moyenne) et cortex auditif secondaire (partie antérieure).

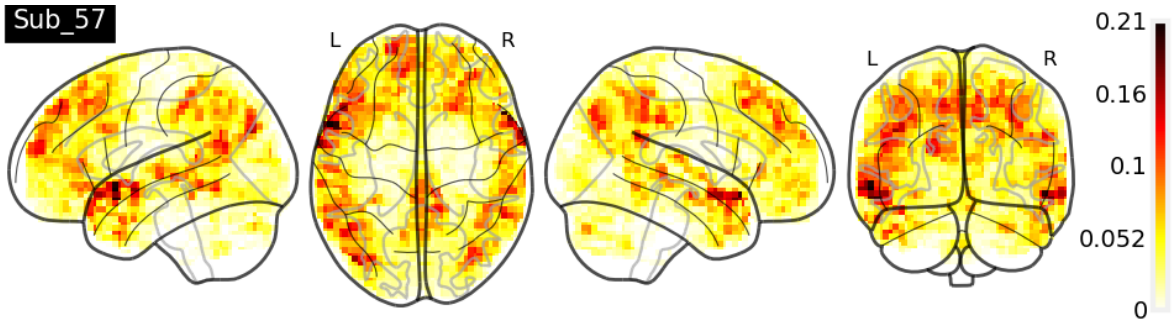


FIGURE 10 – Coefficient de détermination BFWE

Sur les figures 10 et 11, les résultats sont similaires. Une activation au niveau du gyrus angulaire, à la frontière inférieure du lobe pariétal avec le lobe temporal (la jonction temporo-pariétal, TPJ). Au niveau du gyrus frontal inférieur, plus précisément dans l'aire de Broca à la *pars triangularis* il est également possible d'observer un score de prédiction non négligeable. Enfin, le cortex préfrontal médial et une zone temporale, plus postérieure, sont aussi bien prédites.

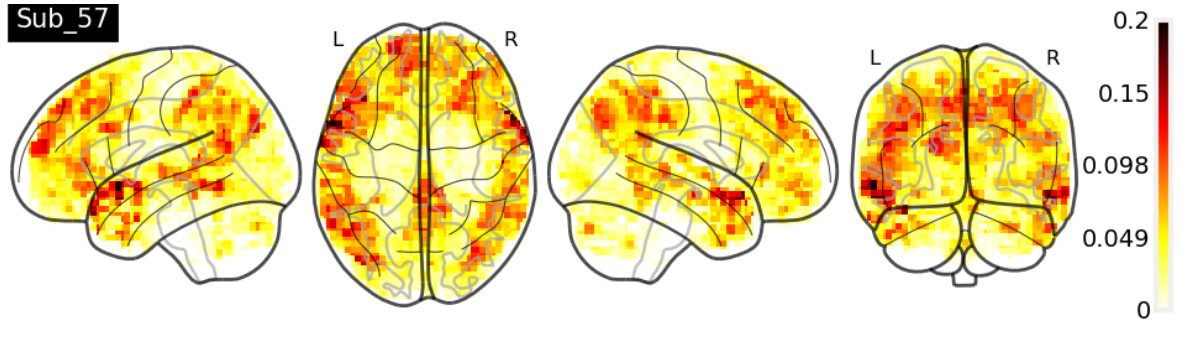


FIGURE 11 – Coefficient de détermination BFLSTM

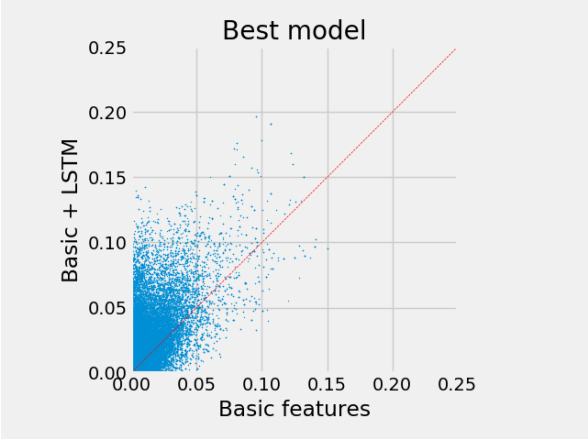


FIGURE 12 – LSTM en fonction de BF

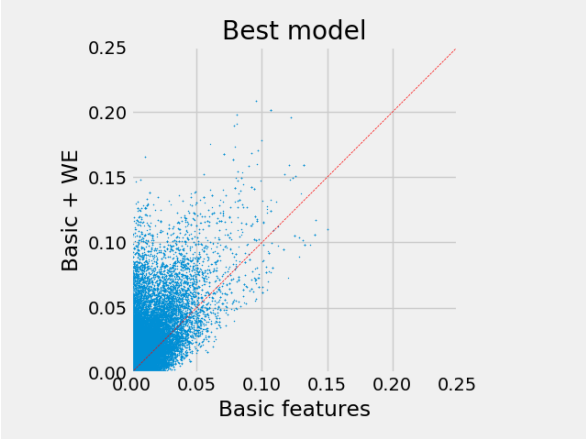


FIGURE 13 – WE en fonction de BF

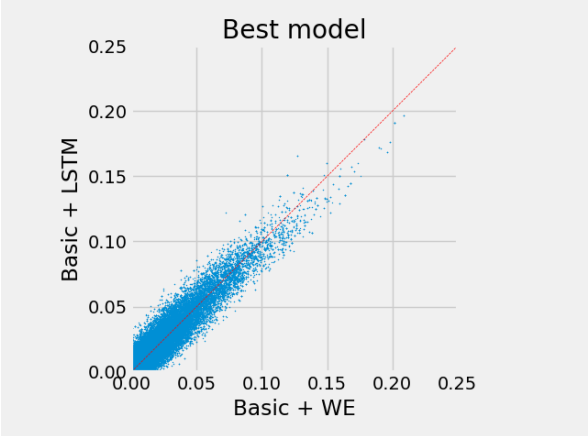


FIGURE 14 – Modèle LSTM en fonction du modèle BF

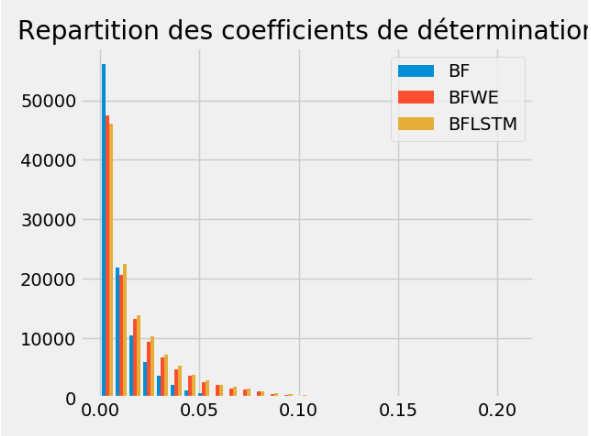


FIGURE 15 – Histogramme de répartition de R^2

4.2 Scatterplot - Histogramme

Les figures 12 à 14 mettent en relation les performances de chaque modèle. Chaque point correspond à un voxel significatif. La coordonnée en abscisse est égale au coefficient de détermination pour ce voxel pour le modèle en abscisse. Parallèlement, l'ordonnée est le coefficient de détermination pour le modèle concurrent. Ainsi, en analysant la distribution de part et d'autre de la diagonale, ces figures fournissent une

indication de la performance relative des modèles. La figure 15 illustre la répartition des valeurs des coefficients de détermination considérant tous les voxels significatifs.

4.3 Optimisation du réseau LSTM

Afin de réduire le temps de calcul et la nombre de *features*, le réseau LSTM a été entrainé sur le corpus de Wikipédia de nombreuses fois selon différents paramètres. L'objectif premier était de trouver une configuration avec un nombre de neurones plus restreint et des performances similaires. Le réseau LSTM ayant déjà été optimisé, cette nouvelle configuration induirait nécessairement une perte de performance du réseau. Cependant, avec moins de *features* la régression par le modèle de prédiction pourrait être plus performante et dépasser l'erreur induite par une moins bonne structure du réseau LSTM. Plusieurs simulations ont ainsi été conduites. La figure 16 illustre les performances du modèle grâce à une valeur de perplexité. En linguistique, la perplexité rend compte de la capacité d'une distribution probabiliste à décrire l'échantillon. Plus la perplexité est faible, plus le modèle est performant.

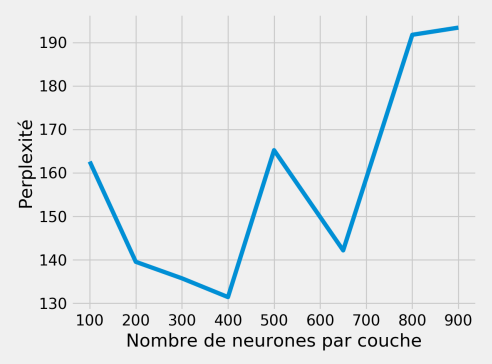


FIGURE 16 – Coefficient de détermination BF

Cette simulation rend compte de certaines configurations qui semblent être autant voir plus performante que la configuration utilisé pour nos expériences.

5 Discussion

5.1 Analyse des résultats

Dans ce mémoire a été exposé une méthode de régression linéaire visant à prédire l'activité cérébrale à partir d'activations HS de réseaux LSTM. Il s'agit d'une preuve de concept montrant que l'information contenue dans ces réseaux n'est pas complètement éloignée de la réalité biologique. La méthode utilisée pour la prédiction sur une régression de Ridge couplée à une double validation croisée hierarchisée. Des résultats de prédiction encourageants ont été obtenus dans les aires propres au langage chez l'homme dans le sujet test.

L'objectif premier de cette étude était de faire ressortir une corrélation entre les réseaux informatiques et biologiques afin de s'instruire sur le fonctionnement des deux parties. De plus, ceci a permis d'objectiver la pertinence de ces modèles à représenter la réalité biologique.

Le modèle BF, principalement construit à partir d'éléments acoustiques et sémantiques a montré des scores de prédiction non négligeables au niveau du cortex auditif primaire (gyrus temporal supérieur, partie moyenne). La présence de *feature* telle que RMS est sans doute déterminante pour cette zone. La partie antérieure du gyrus temporal supérieur, cortex auditif secondaire est aussi concerné. Il reçoit directement les informations venant du cortex auditif primaire et traite les signaux de plus haut niveau.

Le modèle BFWE apparaît plus performant que le modèle BF. L'ajout des Word Embeddings semble réellement entrainer l'activation des zones du langage et plus seulement des zones de l'audition. L'apparition d'activations dans les zones de traitement de niveau divers comme l'aire de Broca *pars triangularis* ou le lobe préfontal médian conforte l'hypothèse que le réseau informatique suit un parcours semblable au cerveau humain dans l'analyse de ses signaux. L'aire de Broca est connue pour être très impliquée dans le traitement du langage et dans la génération de la parole [7].

Le modèle BFLSTM est très similaire au modèle BFWE. Il est intéressant de noter que le réseau LSTM reproduit assez fidèlement les Word Embeddings de lui-même, sans intervention humaine. Ce type de modèle pourrait servir à la génération de Word Embeddings à l'avenir et peut-être même surpasser les techniques existantes en utilisant un algorithme plus sophistiqué. Il serait intéressant d'étudier les différences de distribution des valeurs entre des Word Embeddings généré via GloVE ou via un réseau LSTM.

Les zones les mieux prédites, particulièrement dans les modèles BFWE et BFLSTM sont des zones très importantes pour le langage. En effet, tout comme pour le modèle BF, on conserve les activations dans le cortex auditif primaire suivi du cortex auditif secondaire, ce qui suggère une intégration par le réseau LSTM des caractères acoustiques. Les prédictions au niveau de la jonction temporo-pariétale sont extrêmement intéressantes du point de vue sémantique de la phrase [2]. L'aire de Broca, également visible sur ces figures, s'active physiologiquement lorsqu'une phrase est perçue par l'individu comme correcte [9]. L'aire préfrontale médiale rend compte de fonctions de très haut niveau de cognition. Elle s'active lorsque deux informations présentées semblent cohérentes[1]. Enfin la zone temporale semble être impliquée dans le lexique des mots.

Cette liste importante d'aires cérébrales assez bien prédites par le modèle constitue un résultat encourageant. Le réseau LSTM semble s'être accordé sur des fonctions du langage similaires à celles présentes dans le cerveau humain. L'étude approfondie du modèle peut certainement apporter des éléments supplémentaires sur les zones encodant la sémantique dans le cerveau humain [10].

Les figures 12, 13 et 14 suggèrent que les modèles BFWE et BFLSTM sont plus performant que le modèle BF. On voit en effet la majorité des points au dessus de la diagonale, ce qui signifie que le coefficient de détermination est plus élevée. L'ajout d'information supplémentaires comme les Word Embeddings influe donc positivement sur la performance du modèle. Dans une prochaine étape, il serait intéressant de comparer un modèle LSTM pur (sans les BF) au modèle BF, afin de voir si l'information contenue dans le LSTM est importante.

Enfin, la figure 15 donne un aperçu de la distribution des valeurs du coefficient de détermination. Le modèle BF semble engendrer plus de bruit que les autres modèles de par le fait que beaucoup des voxels significatifs ont un coefficient de détermination bas. Les modèles BFWE et BFLSTM sont eux plus performants et les coefficients de

détermination mieux répartis.

5.2 Limites

L'analyse de ces résultats n'est malheureusement possible pour le moment que sur un sujet. Le temps de calcul d'un modèle peut prendre plusieurs jours. Par optimisation du pré-traitement des données IRMf (moins de voxels, mais plus gros) ainsi que par optimisation des données LSTM (voir Chapitre 4.3) il est possible de diminuer le temps de calcul d'un facteur 8 environ. Cette étude doit se poursuivre et couvrir les 51 sujets du projet. La création d'un modèle plus général par validation croisée entre sujets constitue une prochaine étape.

De plus, la régression de Ridge reste une régression linéaire simple et de meilleurs résultats pourraient être obtenus en implémentant des algorithmes plus sophistiqués au détriment de l'interprétabilité des résultats. Le nombre important de *features* pourrait être ajusté via des algorithmes de sélection de *features*.

L'interprétation des données reste cependant complexe de par la nature saccadée de l'IRMf (un scan toutes les deux secondes, alors que plusieurs mots ont été entendus par le sujet dans cet intervalle de temps).

6 Conclusion

Nous avons ainsi développé un programme de prédiction d'activation IRMf à partir de l'activité d'un réseau LSTM. Les résultats obtenus sont satisfaisants pour une première expérience et les capacités de cet outil sont encore à explorer. Le développement d'un modèle encore plus poussé et performant permettra de nous apprendre beaucoup sur le fonctionnement de notre cerveau. Ce type de modèle informatique reste beaucoup plus simple à manipuler que des systèmes biologiques. Aussi, il est très simple de voir l'évolution de l'information pendant sa traversée du réseau de neurones, ce qui est très complexe *in vivo*.

L'analyse des données françaises permettrait d'en apprendre plus sur les différences structurelles fondamentales d'apprentissages des langages. En parallèle de ces sophistications de la prédiction de base, il paraît intéressant "d'inverser" le processus et de prédire les mots pensés par le sujet en se basant sur des données d'IRMf. Il n'est pas exclu que ce programme inversé puisse prédire un mot proche sémantiquement du mot écouté et ainsi de reconstituer la trame de l'histoire entendue.

En conclusion de ce mémoire, il est primordial de souligner le caractère intriguant et préliminaire des résultats de ce modèle de prédiction. Ce type de modèle faisant partie d'une branche très peu explorée de la neuroimagerie linguistique, il convient de traiter les résultats avec prudence et de poursuivre les investigations.

7 Remerciements

Je remercie chaleureusement le laboratoire Neurospin pour m'avoir accueilli, ce fut un stage très enrichissant et plaisant à vivre. Une énorme reconnaissance à Yair Lakretz et Christophe Pallier qui ont su être extrêmement pédagogues, sympathiques

et présents dès que nécessaire. Enfin, merci à mon ordinateur de ne pas avoir brûlé malgré tous ces calculs.

Références

[1] Marand Raymond A. The neural bases of social cognition and story comprehension. *Annual Rview of Psychology*, 62 :103–134, 2011.

[2] Binder, Jeffrey R., and Rutvik H. Desai. The neurobiology of semantic memory. *Trends in Cognitive Sciences*, 15(11) :527–36, 2011.

[3] Jeffrey R. Binder, Julie A. Frost, Thomas A. Hammeke, Robert W. Cox, Stephen M. Rao, and Thomas Prieto. Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, 17(1) :353–362, jan 1997.

[4] Noam Chomsky. *Syntactic Structures*. 1957.

[5] Allan M Collins and Elizabeth F Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6) :407, 1975.

[6] Allan M Collins and M Ross Quillian. Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior*, 8(2) :240–247, 1969.

[7] Adeen Flinker, Anna Korzeniewska, and Avgusta Y. Shestyuk. Redefining the role of broca's area in speech. *Proceedings of the National Academy of Sciences*, 112(9) :2871–2875, mar 2015.

[8] Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. *The elements of statistical learning*. 2001.

[9] Hickok and Gregory. The functional neuroanatomy of language. *Physics of Life Reviews*, 6(3) :121, 2009.

[10] Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532 :453–458, apr 2016.

[11] Dan Jurafsky and James H. Martin. Speech and language processing, aug 2018. `https://web.stanford.edu/~jurafsky/slp3/`.

[12] Roland Kuhn and Renato De Mori. Cache-based natural language model for speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(6) :570–583, jul 1990.

[13] Kundu, Prantik, Souheil J. Inati, Jennifer W. Evans, Wen-Ming Luh, and Peter A. Bandettin. Differentiating bold and non-bold signals in fmri time series using multi-echo epi. *NeuroImage*, 60(3) :1759–70, 2012.

[14] Nikos K. Logothetis and Brian A. Wandell. Interpreting the bold signal. *Annual Review of Physiology*, 66 :735–769, mar 2004.

[15] Edward E Smith, Edward J Shoben, and Lance J Rips. Structure and process in semantic memory : A featural model for semantic decisions. *Psychological Review*, 81(3) :214, 1974.

[16] Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. Lstm neural networks for language modeling. *INTERSPEECH 2012*, sep 2012.

8 Annexes

Les programmes réalisés pendant ce stage sont en accès libre ici :
https ://github.com/chrplr/lpp-scripts3/tree/master/r2maps-ridge. Le modèle de prédiction se trouve dans le répertoire alignement_model sous le nom de main.py. Un fichier explicatif README.md est présent.